VIVO语义网技术在干部信息管理中的应用

任玮[1]　黄金霞[2]

（1.澳大利亚国立大学工程与计算机学院，堪培拉，澳大利亚；

2.中国科学院文献情报中心，北京，100190，中国）

摘要：干部信息管理系统存在着数据管理效率低下、搜索结果简单的问题。基于VIVO语义网技术在关联数据、本体知识库以及语义搜索的实现机制，设计干部信息智能分析系统架构，构建RDF协议定义的干部信息本体知识库，建立不同类型数据之间的关系映射，实现数据关联和隐性数据发现，提高干部信息管理系统语义搜索效率。结构化和关联性的本体框架也降低信息管理成本。

关键词：VIVO；语义网技术；本体知识库；干部信息管理系统

中图法分类号：TP31　　**文献标识码**：A

VIVO Semantic Web Technology Application on Cadres Information Management

REN Wei[1], HUANG Jin-xia[2]

(1.College of Computer Science and Information Technology , Australian National University, Canberra, Australia;

2.National Science Library, Chinese Academy of Sciences, Beijing 100190,China)

Abstract: Cadre information management systems have some problems such as managing the data inefficiently and providing the search results simply. Basing on VIVO semantic web technologies on linked data, ontology-based knowledge base and semantic search, an intelligence analysis system on cadres information was designed, including defining a semantic search database on RDF protocol, building cadres information ontology-based knowledge base, establishing the relationships between different data types mapping. The system aims to achieve data linked and discover the hidden data, and to improve the efficiency of semantic search on cadre information. Ontology-based system frame on structured and linked date also reduce the management costs on Cadre information.

Key words: VIVO; Semantic web technology; Ontology-based knowledge base; Cadre information management system

1　引言

干部管理信息系统在组织管理和人事管理工作中具有十分核心的地位。随着干部管理的科学化建设，干部资料的信息化，包括干部信息的数字化存储、信息管理、支撑干部遴选决策等，已成为组织部门需要解决的重要问题[1-2]。目前，多数的干部信息管理系统仅仅实现对干部信息的显示和简单检索，无法实现科学管理好干部的信息化要求[3]。迫切需要使用先

收稿日期：修回日期：

进的信息技术，基于较好的资源共享性和智能分析性，建立安全、高效、准确、便捷的干部信息管理流程，让干部信息管理变得科学化、信息化和便捷化，也对干部的选拔任命以及培养等工作起到一定的支撑作用。

当前，国内外搜索引擎依然面临着搜索方法的查全率和查准率不足的危机，虽然谷歌很早就开始研发语义技术，但当前仍旧难以满足特定用户的语义检索需要[4]。语义搜索技术可改善当前搜索引擎的搜索效果，包括查全率和查准率，然而现阶段的语义搜索研究面临着两大难题：一是没有找到一个有效的解析用户语义的手段，二是没有形成一个成熟的语义搜索模型[5]。由于数量容量不同、应用对象不同，目前并不存在一个“通用”的解决方案。以情境理论为理论基础、以本体论为方法论基础，利用本体将抽象的情境因素表达为信息检索系统可以读取和利用的具体变量，将可能发展成为一个以本体为基础的语义信息搜索模型[6]。美国 VIVO 项目采用 RDF、OWL、关联数据等语义网技术，构建了基于本体的语义工具[7]，已经被应用于美国 20 多个高校联合建立的科研网络 VIVO[8]、美国农业部的信息资源发现系统[9]、中国科学院 100 多个研究所的领域知识环境（SKE）[10]等多种用户应用情境中，支持信息的结构化和关联化处理，实现多来源信息的语义搜索。本文针对干部信息科学管理和使用中的问题，尝试利用 VIVO 语义网技术和搜索模型，构建干部信息智能分析平台，以提高了干部信息管理和利用的效率。

2　VIVO 语义网技术

2.1 VIVO 简介

语义搜索将语义 Web 技术引入搜索引擎，改善当前搜索引擎的搜索效果。语义搜索主要可分为基于传统搜索的增强型语义搜索和基于本体推理的知识型语义搜索[5]。基于本体推理的知识型语义搜索，是以本体构建的知识库为主体，通过本体知识库推理实现知识发现型的语义搜索，VIVO 项目成功地体现了其价值。

科学家语义网络（VIVO）最初是美国康奈尔大学图书馆在 2004 年启动，以支撑该校农业和生命科学学院的生命科学研究而发展起来的，在 2007 年利用 RDF、OWL、Jena 和 SPARQL 等技术对 VIVO 进行了改造，用于对康奈尔大学所有院系的教员、科研人员和学科等信息的语义检索和关联发现。基于 VIVO 在康奈尔大学的应用以及科学研究的合作需要，2009 年 9 月份美国 7 所大学联合申请，从美国国立卫生研究院（NIH）的国家研究资源中心（National Center for Research Resources）获得 1,220 万美元的资助，以建设一个开放的语义应用，这个应用是为了寻求一种解决方法，来促进科研人员的科研网络化协作。VIVO 建设的软件和工具是开源的，以便于其他有类似需求的系统建设者能够快速有效地建立应用，实现 VIVO 倡导的构建国家的乃至国际的科学家网络目标[11]。

2.2 VIVO 主要应用技术

VIVO 应用的语义网技术包括关联数据技术和本体技术，这些技术目前已被广泛应用，本文不做赘述。需要说明的是，与其它系统中的应用情况不同，在 VIVO 中这些技术以一种简单明确的方式被用来处理多来源、多类型信息的结构化和关联化，以支持用户对这些数据的语义检索、关联检索，以及系统管理者对底层数据的高效管理。VIVO 项目核心是设计了本体模型和本体知识库框架，为此还开发了本体编辑器-Vitro，用于 VIVO 应用者快速创建本体和知识库，实现检索功能，支持其中 RDF 数据的导入、导出、SPARQL 查询。另外，Vitro 还提供服务显示内容与底层本体知识库的灵活关联，支持 VIVO 应用者快速搭建用户使用平台[12]。

2.2.1 资源描述框架协议和关联数据技术

资源描述框架（Resource Description Framework，简称 RDF）是一个用于表达关于万维网(World Wide Web)上的资源或信息的语言，基本思想是：所有信息都被称为“资源”；被描

述的“资源”具有“属性”，而这些属性各有其“值”；对资源的描述可以通过“陈述”来进行。RDF 采用三元组陈述<主体、谓词、客体>来描述 Web 上的资源[13]。关联数据是一种可行的语义网表达形式，它采用 RDF 数据模型，利用 URI 命名数据实体来发布和部署“类”数据和“实例”数据，从而不同来源可以通过 HTTP 协议揭示并获取这些数据，同时强调数据的相关关联以及有益于人机理解的语境信息。VIVO 利用关联数据技术，实现对分布于各成员学校 VIVO 系统中数据的关联发现，从而创建科学网络。

2.2.2 本体和本体知识库

本体是一种共享的概念化模型的明确且形式化规范说明[14]。共享概念包括用来对领域知识进行建模的概念框架、需要互操作的主体之间用于交互的与内容相关的协议，以及用于表示特定领域的理论的共同约定。知识库是指建立一组结构良好的知识模型并可方便地对相关知识进行描述、整理。利用知识模型可以将知识方便的表示成文本形式，但是文本形式的知识不利于知识的处理，例如推理、分类等操作，基于本体化知识模型的知识库可以支持数据分类、隐性知识推理等[15]。VIVO 本体是 VIVO 建设的核心部分，本体设计策略是构建主从关系的通用核心本体和本地本体，不同命名空间的本体将支持对全部的或本地的本体实例的查询。基于本体模型，外部信息被摄取入 VIVO 进行本体类和属性等的匹配从而实现结构化处理，最终形成实例，进入核心本体知识库或本地本体知识库。生成本体实例的过程，比较复杂，在 VIVO 中是有专门的数据摄取工具，可以参见文献[16]。

3 基于 VIVO 模型的干部信息智能分析平台设计

干部信息智能分析平台是在基于语义搜索这一概念的基础上，开发出针对干部信息管理的功能：通过语义搜索，使计算机能够准确明白用户所输入语句的真正含义，并以此来进行搜索，从而精准的向用户返回最符合的需求，以实现对信息的有效管理。该功能的实现是基于 RDF 协议的本体知识库来进行的干部信息管理和信息搜索，并提供初步的干部信息管理分析能力，如任职经历分析等功能，满足特殊部门的需求。

采用 VIVO 的 Vitro 软件作为平台的构建基础，采用 VIVO 核心本体并构建知识库，定义和描述语词，降低二义性，同时，为提高检索效率，制定新的推理规则和搜索策略，对信息资源进行高效推理。另外，本体的组织结构与当前干部信息管理系统中的传统数据在组织结构上存在巨大差异，因此智能平台与传统平台之间数据转换即如何有效地将传统数据转换成本体知识库所需要的格式，也是智能平台构建时所需要面临的巨大问题。

3.1 系统结构

干部信息智能分析平台具体功能包括提供干部信息的语义搜索、成长经历分析、被搜索人关系网络，以及相关人事管理人员的知识库管理、数据管理以及账户管理等内容。系统结构如图 1 所示。按此功能设计系统模块，主要由两个部分组成，分别为用户应用功能部分、开发和维护功能部分。在用户功能部分中，允许用户进行干部的人名、职务等的语义搜索，查看由其“具体经历”分析的成长历程，发现由规则推理出的该干部人员关系网络；在开发维护部分，支持平台管理人员对信息的管理，包括对干部本体知识库进行创建和修改、数据导入或导出，并实现对不同层级（例如领导、单位内人员、一般访客）的访问权限管理。

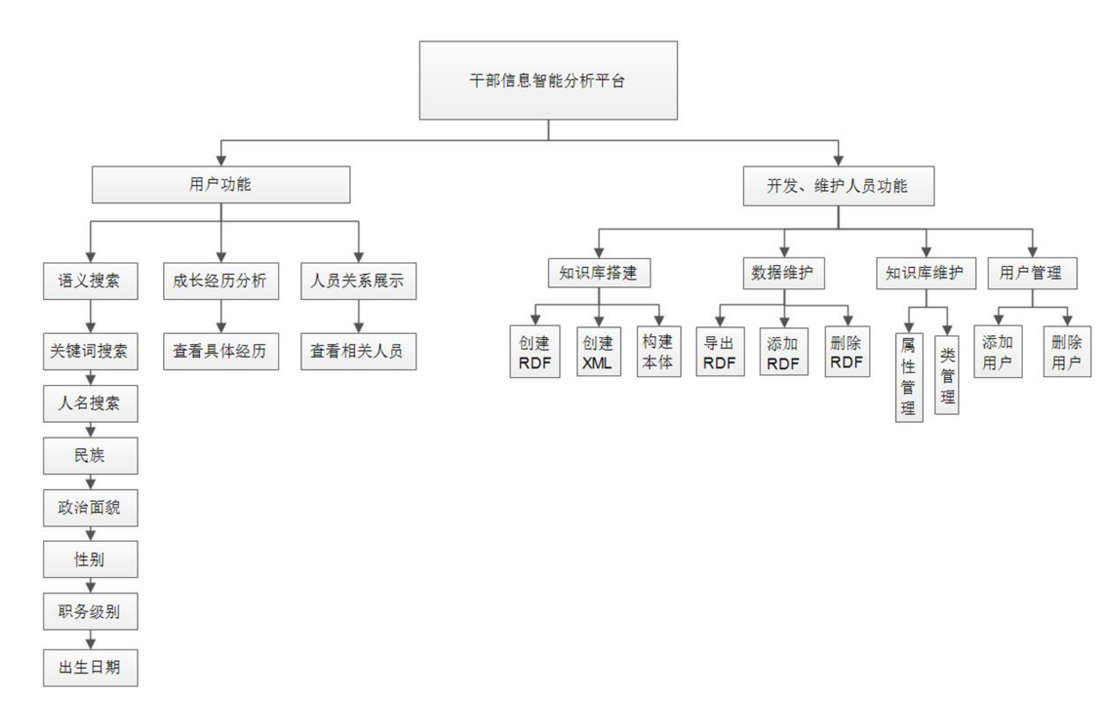


图 1 干部信息智能分析平台系统结构

3.2 关键技术

平台建设过程中遇到的问题包括干部信息本体构建、干部信息本体知识库构建、语义搜索模型构建。

3.2.1 干部信息本体构建

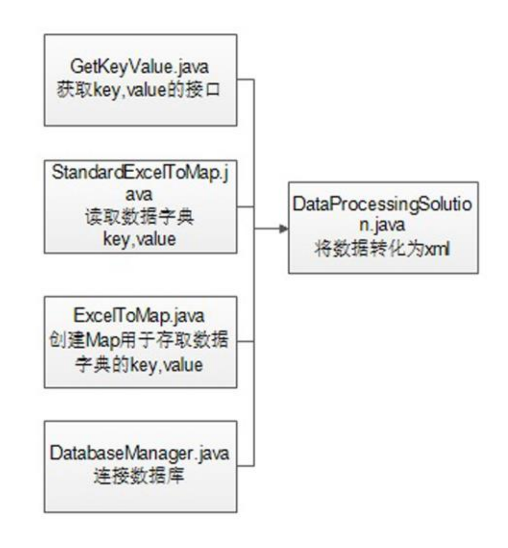
干部信息本体复用了 VIVO 核心本体。VIVO 核心本体定义了 13 个一级类，涵盖了人员、机构、学术活动、成果著作、空间、时间等，并对各类设置了多样的属性，例如"人员"的二级类"管理人员"属性包括姓名、教育经历、工作经历、发表论文、出版著作、参与项目、获得荣誉等[17]。

3.2.2 干部信息本体知识库构建

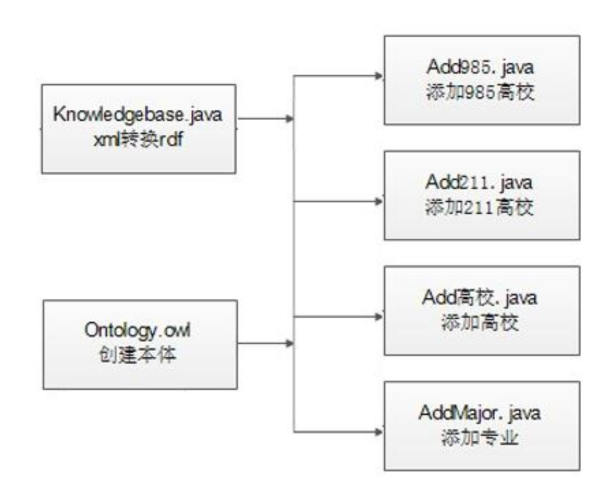
干部信息本体知识库的构建采用 VIVO 提供的应用和相关工具。干部信息智能分析平台知识库中的数据来源，一部分来自原有传统数据库中的数据，一部分来自于平台管理人员或平台用户的自我输入。前者需要系统管理人员操作数据的批量摄取，后者则需要用户按照本体实例进行描述即可。

本体知识库的数据摄取，需要先将传统数据库中的数据按照类别进行划分，实现关系数据到 XML 转换、XML 到 RDF 转换，最终形成本体知识库[18-19]。

第一步：将干部信息原始数据转换成 XML 格式。程序结构如下：



第二步：将 XML 转换成 RDF 形式从而建立完整的知识库。程序结构如下：



3.2.3 语义搜索模型构建

用户在执行检索时，系统将基于关键词的查询处理，提供实例状态的传统搜索结果，同时，利用关联规则和本体中设定的推理规则（即推理引擎），基于检索出的实例，从本体知识库中推理发现相关联的更多实例，即隐性信息的发现。系统中用户查询请求和检索处理方式，用图 2 表示出系统中数据走向。

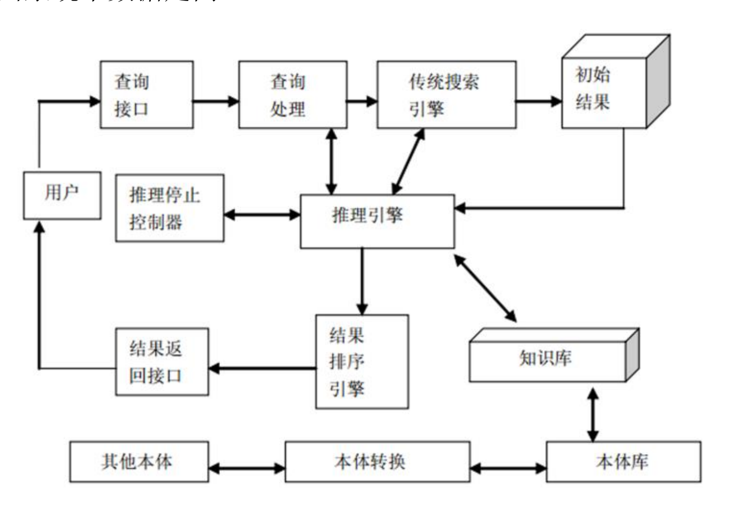


图 2 干部管理智能平台数据走向示意图

3.3 系统实现流程

3.3.1 数据库配置

系统需要从传统数据库或存储数据库中导出基本数据以及数据字典，搭建基本数据库系统是十分必要的。干部信息智能分析平台需要进行复杂的数据转换，在搭建好系统框架后，配置好数据库才可以开始系统的真正实现。

1. 创建 Map 以便获取数据字典

```
public class StandardExcelToMap {
public static Map<String, Map<String,String>> readExcel(File file) throws IOException {
...
```

2. 数据库原始数据向 XML 的转换

```
public class DataProcessiongSolution1 {
...
```

3.3.2 知识库模块实现

知识库的实现是整个系统的基础，这个知识库实现的流程分别是将 Xml 转换成 RDF，同时根据数据库结构创建相对应的本体，最后根据创建的本体构建相应的知识库。

1. 导入所需要的类以及相关推理器，并在 VIVO 空间中命名。

```
public class KnowledgeBase_new {
private static String nameSpace =
"http://www.semanticweb.org/whj/ontologies/2013/4/govermentOgnization#";
```

2. 读取存放在 D 盘中的 XML 文件并创建本体模型

3. 根据数据字典将准备的实例放入 Map 中

```
private static void initialize() {
...
```

4. 根据本体建立干部简历知识库

```
public void addResumeKnowledge(File file) {
...
```

5. 根据本体中的实例创建新的干部简历

```
private void recurCreate(Element el, Individual parentIndividual,long startTime, int level) throws Exception {
...
```

6. 干部简历与本体中的实例进行匹配

```
public Individual entityResolution(OntClass elClass, String entityName) {// 解析出对应实体
...
```

7. 使用 Jena 推理器推理本体

```
Public Map<String, String> queryIndividualByClass(OntClass elClass) {// 根据类名搜所有实体
...
```

8. 将机构实例存入知识库

```
public void addSchoolKnowledge(OntModel model1) {
// 创建OntModel中的知识库，加入一些类
...
```

3.3.3 系统用户管理

根据用户对干部信息管理和应用层次的不同，系统设置普通用户和管理员进行分别管理。普通用户的操作过程是先登录进入系统主界面；执行语义搜索可以进入语义搜索界面，该界面中可以选择关键字搜索和人名搜索，同时可以添加民族、政治面貌、性别、职务级别以及出生日期等内容进行限定，在搜索结果中可以继续对最高学历等内容进行限制修改；执行关系分析，进入关系分析界面，搜索人名后可以显示其周边关系，包括婚姻关系，直系亲

属，旁系亲属，以及近姻亲；执行成长经历分析，可以查看成长经历分析；执行成长地分析、干部评价以及班子配备则可以查看相应内容。管理员登录成功进入主界面，执行知识库维护可以进入知识库维护界面，执行本体列表可以查询现有的本体，执行类层级和累分组则可以对当前数据的类进行管理，执行对象属性层级，数据属性层级以及属性分组则可以查询相印的属性信息，执行编辑 Sparql 和现实 Sparql 则可以对 Sparql 语句进行修改；执行数据管理，进入数据管理界面，执行添加/删除 RDF 数据，可以添加或删除 RDF 数据，执行提取工具，可以下载相应的提取工具，执行 RDF 导出可以导出现有的相关 RDF 数据；执行用户管理，可以进入用户管理界面可以添加、删除或查询某一用户。

4　干部信息智能分析平台功能

4.1 语义搜索功能的实现

搜索模块作为系统前台的核心内容，支撑了用户的主要需求。用户使用搜索条对相关人员进行搜索，系统将显示出常规搜索内容以及语义搜索内容。除了常规搜索内容中展示的用户基本信息资料等，对于语义搜索内容，系统还通过动态表格，图画等形式形象的展示出相关人员的人际关系网络，任职经历以及家庭成员身份等深层次信息。

4.2 关系分析功能的实现

关系分析功能有助于加强对某位干部的了解，是干部信息管理的重点。平台用户可以查看到相关人员的具体人际关系，进而可以更为直接的对其职位变动进行有效的管理。

4.3 成长经历分析功能的实现

干部成长经历分析是该平台的另一重要分析功能。通过直观的图表展示出被搜索人的成长经历，包括：对某人成长经历搜索功能的实现；成长经历分析构成实现；将人员经历进行表格化，以实现最终的可视化操作。

4.4 知识库维护功能的实现

这一功能便于管理员或后期开发人员对知识库的维护，包括：本体的编辑、导入或导出；本体类层级、对象属性、数据属性的控制；数据管理功能的设定，方便系统维护人员对系统数据的维护，包括基本的增删查改；用户管理功能的设定，方便系统维护人员对用户进行管理，包括增删改查用户等功能，也包括用户对自己当前的账户进行管理。

5　结束语

语义搜索服务内容定位在知识密集的领域，其提供的是专业门类系统的、全面的、专深的问题，不是传统搜索引擎的普通用户所需要的服务内容。本文构建的干部信息管理平台是基于本体知识库概念、VIVO语义网以及SPARQL查询语言构建的智能信息分析平台，为当前干部管理中的信息化、科学化要求提供实践思路，这也与当前研究和建设的干部信息管理系统均不同，后者普遍只侧重在系统架构上而不关注对数据的组织、只侧重信息管理而不关注用户使用。基于VIVO提供的技术和工具，干部信息管理平台具有较高的可维护性和可操作性，便于开发和维护，便于管理，减少相关部门在干部信息管理和应用上的经济投入。目前的VIVO工具在关联数据上的价值比较高，若能增强语义丰富工具，将进一步促进信息管理和应用的智能化[20]。

参考文献：

[1] 王映龙,潘瑜春. 基于 ASP.NET MVC 的干部管理信息系统的设计与实现 [D]. 江西农业大学计算机与信息工程学院, 2014.

[2] 刘玮. 高校处级后备干部综合评价信息管理系统的设计初探 [J]. 经营管理者,2015(6):

283.
[3] 杜庆涛. 干部管理系统的设计与实现 [D]. 吉林大学计算机科学与技术学院, 2015.
[4] Amerland D. Google Semantic Search: Search Engine Optimization (SEO) Techniques That Get Your Company More Traffic, Increase Brand Impact, and Amplify Your Online Presence [M]. Que Publishing, 2013.
[5] 文坤梅,卢正鼎,孙小林,等. 语义搜索研究综述 [J]. 计算机科学, 2008,35(5):1-4.
[6] 徐静,孙坦,黄飞燕. 近两年来国外本体应用研究进展 [J]. 图书馆建设,2008(8):84-90.
[7] Katy Borner, et al..VIVO: a semantic approach to scholarly networking and discovery [M]. San Rafael, Calif. Morgan & Claypool, 2012.
[8] VIVO: Research & Expertise Across Cornell [EB/OL]. (2015-11-30) http://vivo.cornell.edu.
[9] VIVO: USDA Science & Collaboration [EB/OL]. (2015-11-30) http://vivo.usda.gov.
[10] 宋文,黄金霞,刘毅,汤怡洁.面向知识发现的 SKE 关键技术及服务 [J]. 现代图书情报技术, 2012,7/8:13-18.
[11] Krafft D B, Cappadon N A, Devare B M, et al.. VIVO: Enabling National Networking of Scientists [EB/OL]. (2015-11-30) https://www.researchgate.net/publication/228564370_VIVO_Enabling_National_Networking_of_Scientists.
[12] 黄金霞,宋文,刘毅,等. 中国科学院专业领域知识环境的建设与应用[J]. 图书情报工作(增刊 2), 2010:337-341,352.
[13] 佟强,张富,程经纬,等. RDF 在关系数据库中的存储研究 [J]. 东北大学学报(自然科学版), 2015,36(3): 346-349.
[14] Gruber T..Toward principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human and Computer Studies, 43(5/6):907-928.
[15] 袁磊,张浩,陈静,等. 基于本体化知识模型的知识库构建模式研究 [J]. 计算机工程与应用,2006(30):65-68,104.
[16] 黄金霞,景丽. 面向 VIVO 本体的数据摄取工具[J]. 现代图书情报技术,2011,22(1):16-20.
[17] Mitchell S, Chen S S, Ahmed M, et al..The VIVO Ontology: Enabling Networking of Scientists [EB/OL]. (2015-22-30) https://www.researchgate.net/publication/228934646_The_VIVO_Ontology_Enabling_Networking_of_Scientists.
[18] 李文.Web 信息提取技术研究与应用分析 [J]. 电子技术与软件工程, 2015(3):15.
[19] 何少鹏,黎建辉,沈志宏,等.大规模的 RDF 数据存储技术综述[J]. 网络新媒体技术, 2013,2(1):8-16.
[20] 黄金霞.支持科研和学术发现的语义网应用实例研究 [J]. 图书情报工作, 2011,55(6):126-129.

作者简介：
任玮（1991-），男，北京市人，硕士研究生在读，澳大利亚国立大学工程与计算机学院，主要研究方向为数据分析与语义搜索;
黄金霞（1972-），女，安徽天长人，博士，硕士研究生导师，副研究馆员，主要研究方向为领域本体构建、开放科技资源的数据利用等。